Metabarcoding and Metagenomics 5: 207–217
DOI 10.3897/mbmg.5.67862

Data Paper

Revision and annotation of DNA barcode records for marine invertebrates: report of the 8th iBOL conference hackathon

Adriana E. Radulovici[1*], Pedro E. Vieira[2,3], Sofia Duarte[2,3], Marcos A. L. Teixeira[2,3], Luisa M. S. Borges[4], Bruce E. Deagle[5], Sanna Majaneva[6,7], Niamh Redmond[8], Jessica A. Schultz[1,9], Filipe O. Costa[2,3]

1 *Centre for Biodiversity Genomics, University of Guelph, Guelph, Canada*
2 *Centre of Molecular and Environmental Biology (CBMA), University of Minho, Braga, Portugal*
3 *Institute of Science and Innovation for Bio-Sustainability (IB-S), University of Minho, Braga, Portugal*
4 *L³ Scientific Solutions, Geesthacht, Germany*
5 *Australian National Fish Collection, Commonwealth Scientific and Industrial Research Organisation, Battery Point, Tasmania, Australia*
6 *Department of Biology, Norwegian University of Science and Technology, Trondheim, Norway*
7 *Department of Arctic and marine biology, UiT The Arctic University of Norway, Tromsø, Norway*
8 *Smithsonian Institution Barcode Network, Smithsonian National Museum of Natural History, Washington, DC, USA*
9 *Department of Integrative Biology, University of Guelph, Guelph, Canada*

Corresponding authors: Adriana E. Radulovici (aradulov@uoguelph.ca), Filipe O. Costa (fcosta@bio.uminho.pt)

Academic editor: Fedor Čiampor Jr | **Received** 4 May 2021 | **Accepted** 8 December 2021 | **Published** 16 December 2021

Abstract

The accuracy of specimen identification through DNA barcoding and metabarcoding relies on reference libraries containing records with reliable taxonomy and sequence quality. The considerable growth in barcode data requires stringent data curation, especially in taxonomically difficult groups such as marine invertebrates. A major effort in curating marine barcode data in the Barcode of Life Data Systems (BOLD) was undertaken during the 8th International Barcode of Life Conference (Trondheim, Norway, 2019). Major taxonomic groups (crustaceans, echinoderms, molluscs, and polychaetes) were reviewed to identify those which had disagreement between Linnaean names and Barcode Index Numbers (BINs). The records with disagreement were annotated with four tags: a) MIS-ID (misidentified, mislabeled, or contaminated records), b) AMBIG (ambiguous records unresolved with the existing data), c) COMPLEX (species names occurring in multiple BINs), and d) SHARE (barcodes shared between species). A total of 83,712 specimen records corresponding to 7,576 species were reviewed and 39% of the species were tagged (7% MIS-ID, 17% AMBIG, 14% COMPLEX, and 1% SHARE). High percentages (>50%) of AMBIG tags were recorded in gastropods, whereas COMPLEX tags dominated in crustaceans and polychaetes. The high proportion of tagged species reflects either flaws in the barcoding workflow (e.g., misidentification, cross-contamination) or taxonomic difficulties (e.g., synonyms, undescribed species). Although data curation is essential for barcode applications, such manual attempts to examine large datasets are unsustainable and automated solutions are extremely desirable.

Key Words

annotation, data curation, DNA barcoding, marine invertebrates, metabarcoding, reference libraries

Introduction

Reference libraries, which are collections of compliant DNA sequences assigned to species, constitute the backbone of species identification systems based on DNA barcoding and metabarcoding, and therefore, a critical component in molecular biomonitoring and molecular taxonomy (Weigand et al. 2019). The number of DNA sequences and species included in reference libraries has increased dramatically over the last 15 years (Porter

* Current address: McGill University, Montreal, Canada (adriana.radulovici@mcgill.ca).

Copyright *Adriana E. Radulovici et al.*. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

and Hajibabaei 2018). To date, the ever-growing libraries have been deposited mostly in two large and public molecular databases, namely (i) GenBank (Sayers et al. 2021), a repository with data usually released after publication, and (ii) the Barcode of Life Data Systems (BOLD, Ratnasingham and Hebert 2007), a workbench in which data can be validated and analyzed before being released. Additional databases do exist, but they are smaller in size and usually created for specific purposes (e.g., zooplankton identification, Bucklin et al. 2021).

Along with this expansion, reports of inaccurate or discordant data have become more common (Meiklejohn et al. 2019, Ramirez et al. 2020, Fontes et al. 2021), especially when comparing data of morphological and molecular origin. This is recognized as a critical concern for the accuracy of existing and future DNA-based biomonitoring (Leese et al. 2018, Grant et al. 2021). Whereas any data discordances or inaccuracies in reference libraries would be more apparent using the "conventional" low-throughput DNA barcoding, they can be easily overlooked when applying high-throughput metabarcoding, where automated bioinformatics tools assign taxonomy to millions of sequences (Cristescu 2014). Because taxonomic assignments for metabarcoding data often do not undergo further scrutiny, any faulty assignments due to inaccurate reference library data can be repeated over and over without being noticed (Leese et al. 2016). Since conflicting findings cannot typically be controlled by algorithms, errors in the reference libraries for a particular taxon (i.e., incorrect sequence assigned to a species) might invalidate genuine sequences. The long-term consequences of this erroneous data for biomonitoring and ecological evaluation might be considerable, especially with the planned installation of semi-automated and large-scale metabarcoding-based biomonitoring systems (e.g., BIOSCAN, Hobern 2021).

Data discordances in reference libraries have multiple origins that can be split into two broad categories. First, there are discordances that are due to real biological complexities. Some of these are merely a reflection of the inherent uncertainties and dynamics of alpha taxonomy (e.g., Padial and De La Riva 2006, 2020). Others result from a mismatch between morphological and molecular diagnosis of species boundaries, that is, species that can be identified morphologically but not through short DNA sequences, or vice versa, for intrinsic reasons. The second category of discordances are those introduced through operational errors. For instance, morphology-based specimen misidentifications have been often reported as a source of error (Lis et al. 2016, Pentinsaari et al. 2020). The experience level of the identifier may play a role in such errors, but often taxonomic keys, drawings and descriptions are incomplete or have poor quality, contributing to misidentifications. This may occur in the case of some species that are frequently, but incorrectly, reported as cosmopolitan species across the world because the original description is imprecise enough to include a variety of other species that remain undetected (Gómez et al. 2007, Padial and De la Riva 2020). Taxonomic variants such as synonyms and alternate representation designating the same taxon, are an additional source of mismatches (e.g., *Magallana gigas* and its alternate representation, but widely used name, *Crassostrea gigas*, Salvi et al. 2014, Bayne et al. 2017, Backeljau 2018). These types of inaccuracies and limitations are customarily shared and experienced by biodiversity databases (Bidartondo 2008, Patterson et al. 2010, Meiklejohn et al. 2019). Data discordances due to operational errors are also known to arise during collection, sampling and laboratory procedures, such as specimen and/or tissue sample mislabeling, cross-contamination, or non-targeted PCR amplification (Buhay 2009, Siddall et al. 2009, Evans and Paulay 2012). Poor sequence quality, sequencing errors or usage of reverse strand sequences may also contribute to discordances (reviewed by Pentinsaari et al. 2020).

A workbench such as BOLD, where data can be easily corrected if needed, brings great value to the barcoding and metabarcoding pipelines. Several tools for automated data quality control have been implemented in BOLD, including flags to indicate if sequences of barcode markers (COI, MatK, RbcL, RbcLa, trnH-psbA, ITS, ITS2) are barcode compliant or if the protein-coding genes include stop-codons or common contaminants (e.g., human, cow, mouse, pig, bacteria). In addition, several analytical tools allow data congruence verification. For instance, discordances between species names attributed by BOLD users and the Barcode Index Numbers (BINs, Ratnasingham and Hebert 2013), which are automatically assigned to COI sequences uploaded in BOLD, can be easily revealed through the BIN discordance report. While automated bioinformatic procedures can readily include flags and verification tools (e.g., Andújar et al. 2021), some inconsistencies will eventually require human-mediated inspection and judgement. In fact, if a potential misidentification or contamination is detected, the BOLD team review data and flag records for exclusion from the identification engine (BOLD IDS). The sheer volume and diversity of data to be handled, on the other hand, preclude a complete examination and necessitate a more structured and feasible approach.

Reference libraries have been populated in part through dispersed contributions, despite a few central core facilities providing major inputs (e.g., Canadian Centre for DNA Barcoding, https://ccdb.ca). As a result, DNA sequence data, and respective metadata, which are uploaded to genetic data repositories such as BOLD or GenBank, have varied components and levels of compliance. The research practice also differs among target taxonomic groups, affecting even the type of vouchering system and metadata typically collected and accompanying each specimen (Rimet et al. 2021). The diversity of contributors and the peculiarities of their research practices create chances for operational discordances and shortcomings, resulting in greater difficulties in reference library revision and curation.

The rationale for reviewing barcode data for marine invertebrates is particularly relevant. Marine invertebrates

are often studied as a community and are one of the customary targets for marine biomonitoring using metabarcoding (Duarte et al. 2021). Some taxonomic groups or geographical regions have reference libraries that are far from being comprehensive (McGee et al. 2019, Leite et al. 2020). For instance, only 22%–48% of European marine species, depending on the species list used, had at least one barcode in BOLD in 2019 (Weigand et al. 2019). Furthermore, it was estimated that a robust population of DNA reference libraries may require even decades in some cases (Vieira et al. 2021). Beyond these difficulties, data inaccuracies or discordances in marine invertebrates might be caused by the more incipient status of the overall taxonomic knowledge and various other difficulties including the taxon-specific research practice and reduced number of available taxonomic experts, compared to other more well-studied groups such as vertebrates and insects (Radulovici et al. 2010, Mammola et al. 2020). As a result, a complete human-mediated review and annotation of barcode data might be extremely beneficial for marine invertebrates.

To accomplish this ambitious goal, of manually curating marine invertebrate barcode data, a hackathon was organized in the scope of the 8th International Barcode of Life (iBOL) conference (Trondheim, Norway, 2019). A group of researchers involved in marine invertebrate barcoding were convened with the purpose of undertaking a comprehensive review and annotation of the barcode records of the most representative marine invertebrate taxa currently available in BOLD. The choice to focus exclusively on this platform was based on it being the largest database designed primarily for DNA barcodes and their metadata, the existence of analytical tools embedded in the platform, and the routine process of mining data from GenBank into BOLD, thus ensuring that all DNA barcodes are hosted in one place and circumventing the preference of various researchers for different data repositories. This is a report on the approach, findings and implications for issues related to the curation of reference libraries of DNA barcodes.

Methods

BOLD is a global database structured by few mandatory fields (e.g., phylum, country of collection, and institution storing voucher specimens), including habitat as an optional field overlooked in many records, therefore a specific workflow (Fig. 1) was developed to consolidate only marine data for curation purposes. A copy of the World Register of Marine Species (WoRMS), the most comprehensive species list for the marine realm, was downloaded on June 1, 2019. Only accepted species names for extant marine animals were retained, and further filtering was performed to reduce the list to those invertebrate taxa that are mostly used in marine biomonitoring: crustaceans, echinoderms, molluscs, and polychaetes. For practical reasons to facilitate the curation process during the hackathon, only limited groups were selected in the final list: Crustacea (Malacostraca: Amphipoda, Isopoda, Mysida, and Euphausiacea), Echinodermata (all classes), Mollusca (Bivalvia and Gastropoda) and Polychaeta (all orders). The resulting subset of species from WoRMS was then compared against BOLD, and matched species names had their public BOLD records added to pre-made BOLD datasets. Initially split by taxonomy (phylum), some of the larger datasets (e.g., molluscs) were further divided to reduce the number of records reviewed by one person during the hackathon. Only records assigned to a BIN (Barcode Index Number) were retained in order to use BIN-based analytical tools available in BOLD. BINs are persistent clusters generated periodically and algorithmically for COI sequences with certain quality standards (<1% ambiguities, > 500 base pairs in length) and can be visualized through individual web pages. All public records within each BIN were included in the review. For each dataset, a BIN discordance report was generated, as well as a neighbour-joining tree (NJ tree) with matching specimen details and images, if available, for each sequenced organism.

The revision workflow (Fig. 1) consisted of manually inspecting the discordant BINs (i.e., one BIN with records bearing different species names) together with the NJ tree, followed by searching molecular databases (GenBank and BOLD) and published articles to gather additional information. Those records deemed uncertain, based on the investigation conducted, were annotated with one of four pre-established tags:

a) MIS-ID (misidentification or contamination) – records believed to be misidentified, mislabeled or contaminated,
b) AMBIG (ambiguous) – records that could not be resolved with the existing data,
c) COMPLEX (species complex) – records belonging to species with multiple BINs and, therefore, indicative of hidden or undescribed diversity,
d) SHARE (shared barcodes) – records belonging to species known to be sharing barcodes, due to incomplete lineage sorting or hybridization, based on existing literature.

Each uncertain record was annotated with only one tag. MIS-ID tags were considered the most important since all unflagged records are used for BOLD IDS, therefore they took precedence in cases where one record was falling under multiple tags (e.g., MIS-ID and COMPLEX). Since tags were applied to records and species were usually represented by multiple records, it follows that while any given record can have only one tag, each species may have multiple tags.

The hackathon included only the inspection of COI sequences and not the inspection of morphological specimens stored around the world, resulting in a small degree of uncertainty related to the general findings. For instance, if a BIN included dozens to hundreds of sequences

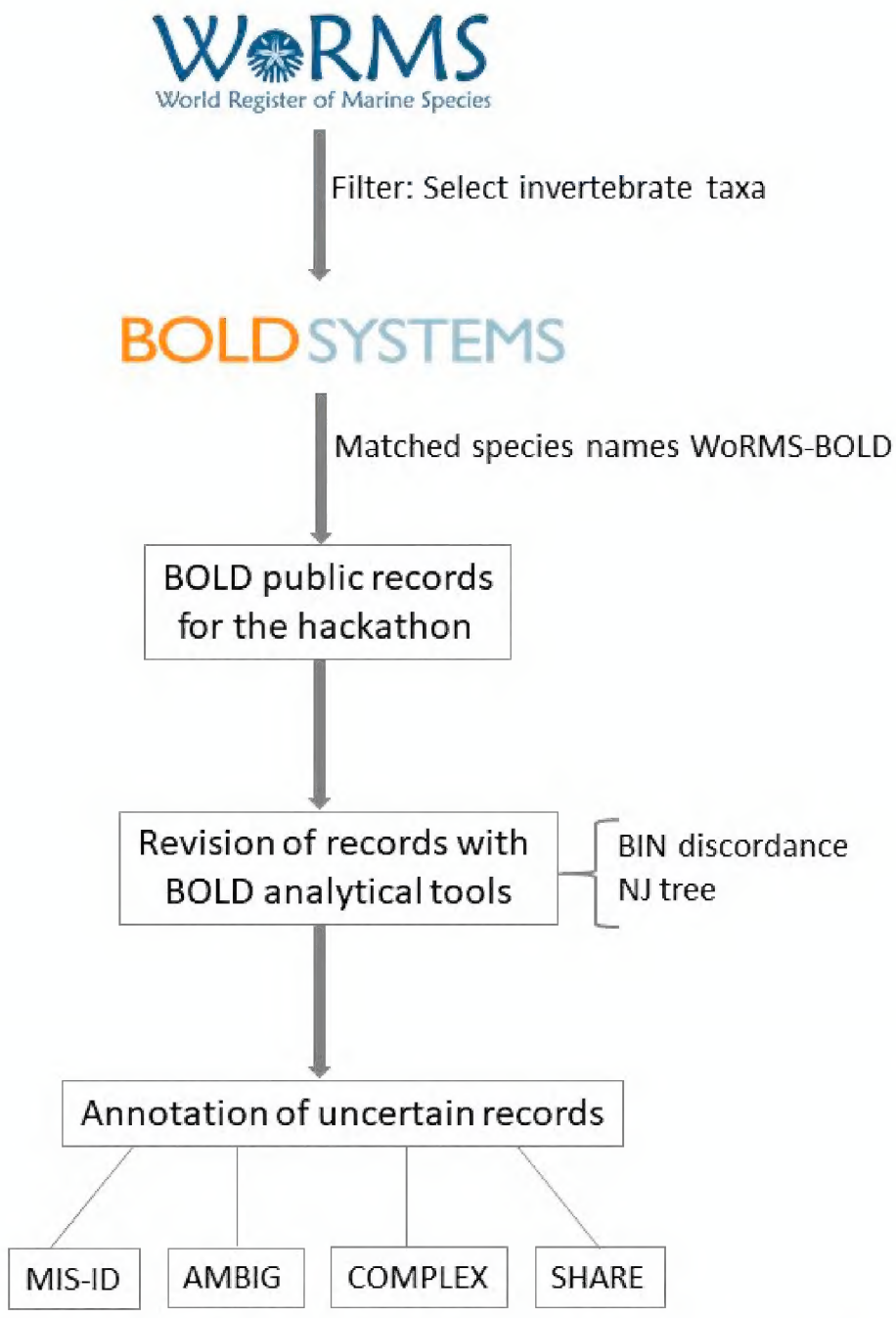

Figure 1. Workflow employed for the review and annotation of selected marine invertebrate records in BOLD. A subset of targeted invertebrate taxa was created from the initial list downloaded from WoRMS. This list was cross-referenced with the available taxonomic list from BOLD. Subsequently, only public BOLD records assigned to a BIN were integrated in datasets and screened with two analytical tools (BIN discordance report and neighbour-joining (NJ) tree). Records deemed to be uncertain were annotated with four pre-established tags: MIS-ID (misidentification, mislabeling or contamination), AMBIG (ambiguous record), COMPLEX (species complex), SHARE (barcode sharing between species). Records suspected to be misidentified or contaminated were annotated and subsequently removed by the BOLD team from the BOLD identification engine (BOLD IDS). Records deemed reliable were not annotated.

of species A and one sequence of species B, the record of species B was tagged as MIS-ID although other possibilities are also viable (species B is correct and species A is incorrect; they are both incorrect; they are both correct, in case of unknown shared barcodes). All the records tagged as MIS-ID were submitted to the BOLD team so they can also be flagged and removed from the database used for BOLD IDS. BOLD allows all its users to insert tags as a tool for data curation by the barcoding community. In contrast, flags can only be added by the BOLD team since they affect BOLD IDS. All flags and tags can be removed by the BOLD team, if necessary.

While detailed attention was given to discordant BINs, records in concordant BINs (i.e., BINs including records bearing only one species name) and singleton BINs (i.e., BINs represented by only one record) were also reviewed, especially in cases of species with multiple concordant BINs (COMPLEX tag). Singletons were not annotated unless they were part of a species complex. The review of records (i.e., assignment of tags as well as additional notes) was recorded directly in the spreadsheets generated by BOLD as matching files for the NJ trees. Formulas were inserted to summarize the findings (number of records tagged, number of records and species per tag type, and number of taxa reviewed at each taxonomic rank). Due to the large amount of data requiring verification and the short time available, the work initiated during the hackathon continued during the months following the event. The results were illustrated through bar graphs using GraphPad Prism 9.0 (San Diego, CA, USA).

All the records reviewed can be found in BOLD (dx.doi.org/10.5883/DS-HACK2019 and dx.doi.org/10.5883/DS-MOLL2019), and all the files with annotations are available in the Suppl. material 1: Tables S1–S10.

Results

The initial WoRMS download had over 600,000 names from all taxonomic levels, but only approximately 200,000 names were accepted animal species names. Further filtering to invertebrate taxa of interest reduced the species list to 79,251 names as follows: Crustacea – 15,148 species, Echinodermata – 7,404 species, Mollusca – 44,883 species, and Polychaeta – 11,816 species. Only a small percentage of these species (about 10%) had barcode representation in BOLD (Table 1).

Globally, the hackathon effort resulted in the review of 83,712 DNA barcode records, distributed across 8,465 BINs, corresponding to 7,576 marine invertebrate species from four phyla, 115 orders, 595 families and 2,490 genera (Table 1). Mollusca was by far the largest phylum tackled during the hackathon (around 50,000 records), thus it was

Table 1. Distribution of the reviewed DNA barcode records among the major taxonomic groups, taxonomic ranks and BINs analyzed, together with the number of tagged (MIS-ID, AMBIG, COMPLEX, SHARE) DNA barcode records and species.

Taxonomic Group	Phyla	Orders	Families	Genera	Species	BINs	DNA barcode records	Tagged species	Tagged records
Bivalvia	Mollusca	26	71	279	741	672	10,194	330	5,088
Gastropoda	Mollusca	38	233	1,066	3,982	4,235	39,749	1,582	15,581
Crustacea	Arthropoda	5	107	349	828	1,129	12,647	290	6,443
Echinodermata	Echinodermata	34	123	447	1,053	1,228	12,756	390	6,155
Polychaeta	Annelida	12	61	349	972	1201	8,366	365	3,434
Total	**4**	**115**	**595**	**2,490**	**7,576**	**8,465**	**83,712**	**2,957**	**36,701**

split into two separate groups, Bivalvia and Gastropoda, for all the subsequent analyses presented here.

Gastropoda was the taxonomic group with the highest number of reviewed records (47.5%) and the highest number of species (53%) in the dataset (Table 1, Fig. 2). On the other hand, Polychaeta was the taxonomic group with the lowest number of reviewed records (ca. 10%), whereas Bivalvia was the group displaying the lowest number of species, comprising about 10% of the total number of species in the dataset (Table 1, Fig. 2).

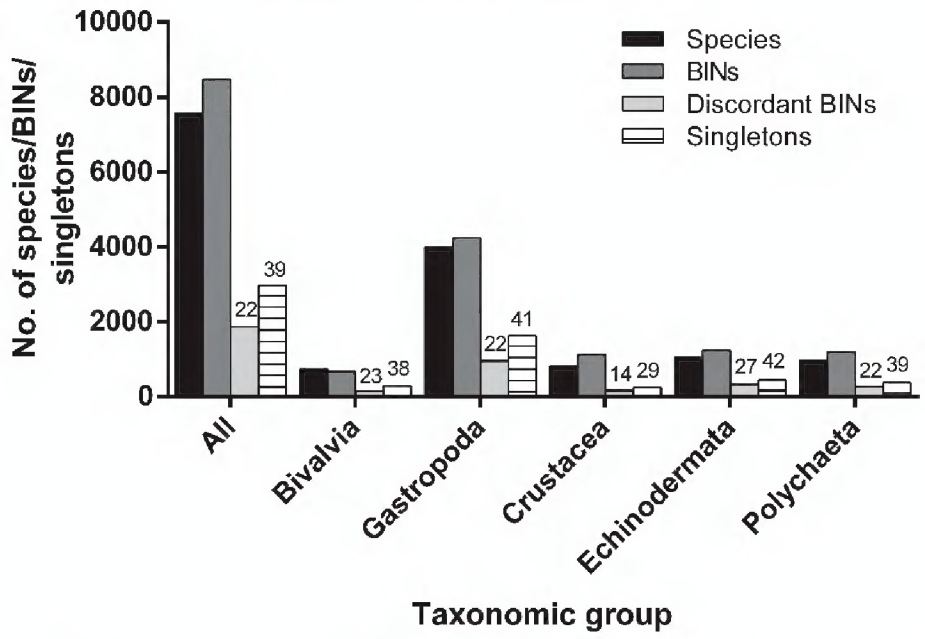

Figure 2. Number of species, BINs, discordant BINs, and singletons (species with only one DNA barcode record) for all groups analyzed and for each major taxonomic group separately. Numbers above bars indicate the percentage of discordant BINs and singletons, respectively.

The number of BINs was highest in Gastropoda and lowest in Bivalvia (Table 1, Fig. 2), and the ratio of BINs/species was above 1.0 in all groups, except Bivalvia (0.9). This indicates that the occurrence of species displaying multiple BINs was prevalent in most groups (Table 1, Fig. 2).

Across the entire dataset reviewed, approximately 22% of BINs displayed discordance (Fig. 2). The highest number of discordant BINs was found in Gastropoda, and the lowest in Bivalvia. Echinodermata was the group displaying the highest discordance in their BINs (ca 27%) and Crustacea the lowest (ca 14%) (Fig. 2). Approximately 39% of the total number of analyzed species were singletons, i.e., represented by a single DNA barcode record on BOLD (Fig. 2). Echinodermata and Gastropoda were the taxonomic groups with the highest percentage of singletons (41 and 42%, respectively), whereas the lowest was recorded in Crustacea (29%) (Fig. 2).

Nearly 39% of all species in the dataset were deemed uncertain (Fig. 3) and tagged with one of the four initially defined tags: MIS-ID (7%), AMBIG (17%), COMPLEX (14%) and SHARE (1%). Gastropoda had the highest proportion of tagged species (ca 54%) and Crustacea the lowest (ca 10%). Globally, the majority of the species were tagged with AMBIG (ca 44%), and the value ranged from ca 30% to 50%, for Crustacea and Gastropoda, respectively (Fig. 4A). About 35% of the tagged species were assigned a COMPLEX tag, with a higher percentage found among Polychaeta (59%), Crustacea (ca 58%) and Echinodermata (43%) and the lowest within Bivalvia and Gastropoda (ca 25% for each class) (Fig. 4A). On the other hand, only about 18% and 2% of the tagged species were classified as MIS-ID and SHARE, respectively (Fig. 4A). The highest proportion of the MIS-ID tag was recorded in Bivalvia (ca 29%), and the lowest in Polychaeta (9%). For the SHARE tag, the highest percentage was observed in Gastropoda (ca 4%), whereas Echinodermata and Polychaeta had no species tagged with SHARE (Fig. 4A).

Approximately 44% of all reviewed DNA barcode records were tagged with one of the four initially defined tags: MIS-ID (3%), AMBIG (10%), COMPLEX (29%) and SHARE (2%) (Table 1, Fig. 4B), with Gastropoda displaying the highest proportion of tagged records (ca 42%) and Polychaeta the lowest (ca 9%). Globally, most records were tagged with COMPLEX (ca 66%), and the value ranged between ca 50% and 92% for Gastropoda and Crustacea, respectively (Fig. 4B). About 23% of the total tagged records were AMBIG, with the highest proportion observed in Gastropoda (ca 33%) and Bivalvia (25%) and the lowest in Crustacea (ca 3%) (Fig. 4B). On the other hand, only about 7% and 4% of the tagged records were classified as MIS-ID and SHARE, respectively (Fig. 4B). The highest percentage of the MIS-ID tag was found in Bivalvia (ca 20%), and the lowest in

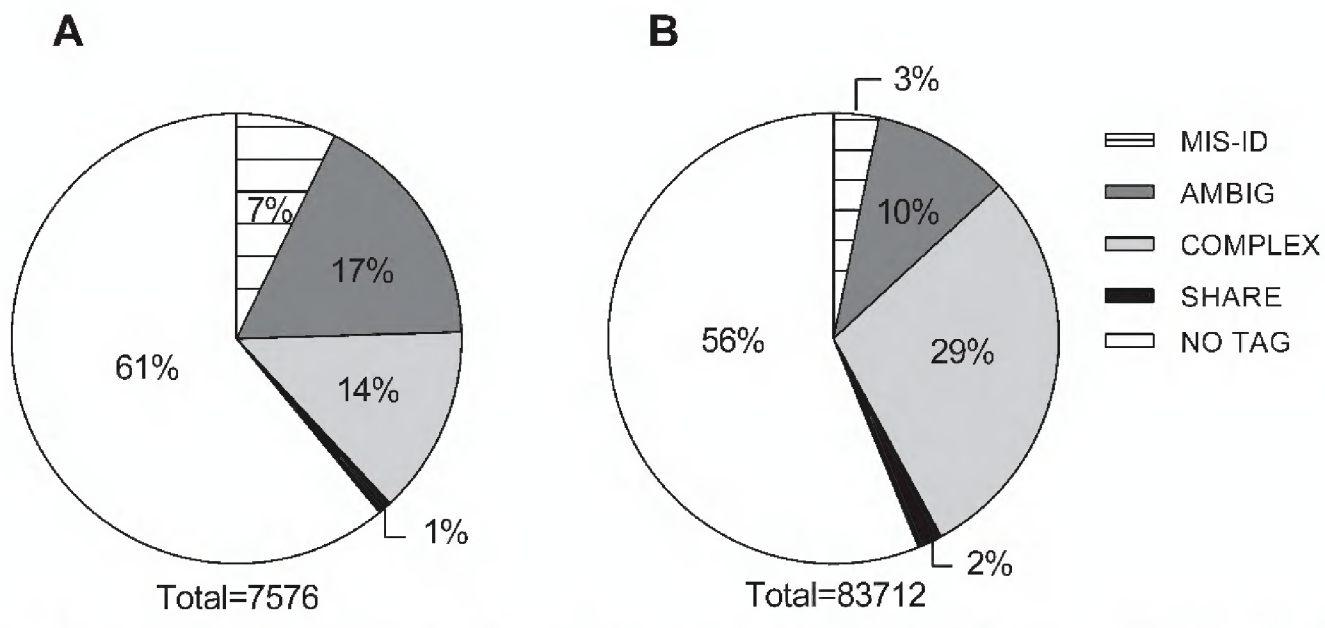

Figure 3. Distribution of the proportion of different tags in the reviewed dataset, in terms of species (A) and DNA barcode records (B). The total number of species (A) and records (B) are added below the chart.

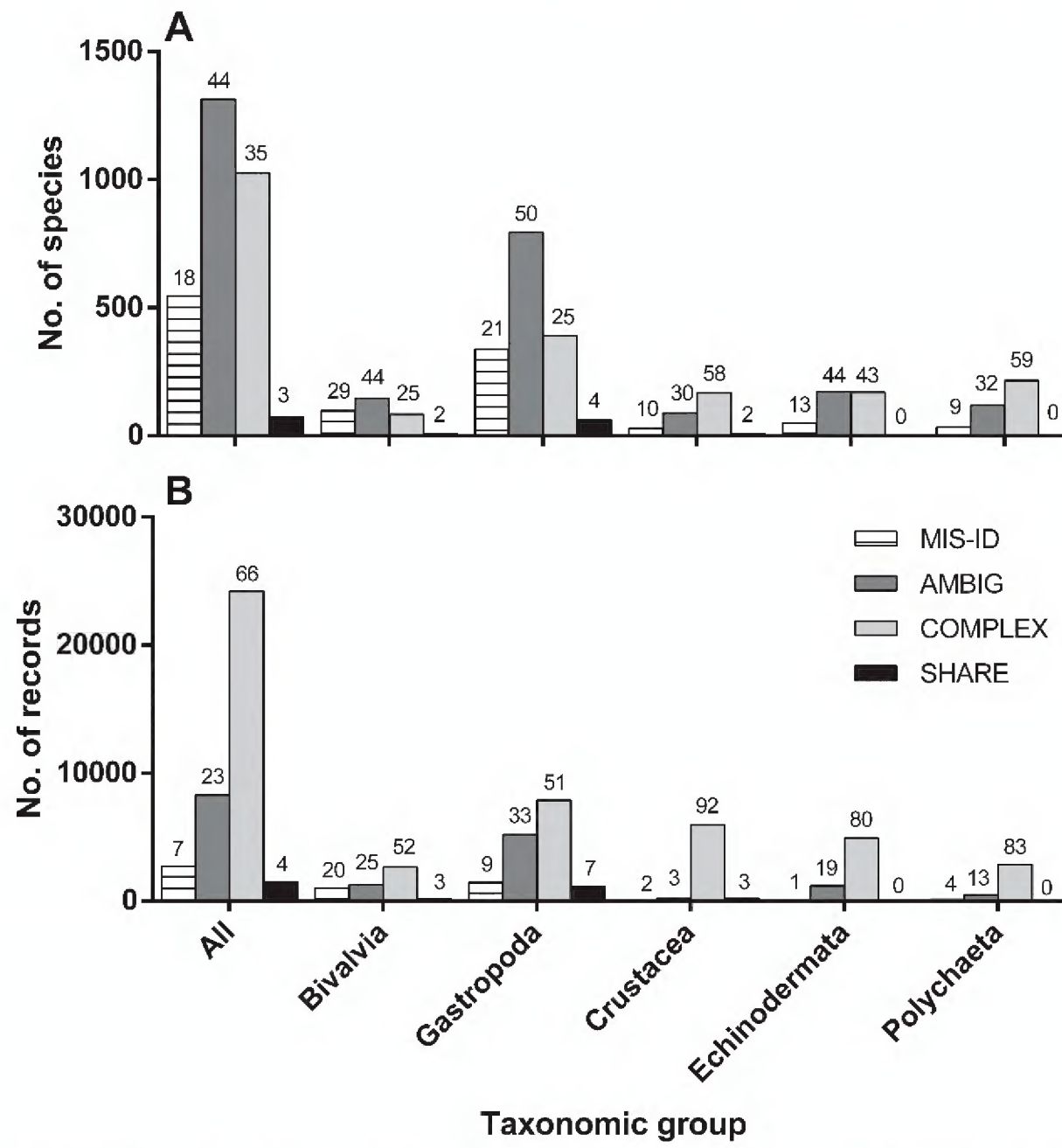

Figure 4. Distribution of the MIS-ID, AMBIG, COMPLEX, and SHARE tags among the major taxonomic groups, considering either the total number of species (A) or the total number of DNA barcode records (B). Numbers above bars indicate the percentage recorded within the whole tagged dataset and within each major taxonomic group.

echinoderms (1%). Concerning the SHARE tag, the highest proportion was recorded in Gastropoda (ca 7%), and no records were tagged with SHARE in Echinodermata or Polychaeta (Fig. 4B).

Discussion

The hackathon on marine invertebrate barcodes fulfilled a variety of purposes beyond the immediate verification of the congruence between morphology and molecular data, and subsequent revision and annotation of records submitted to BOLD. To our knowledge, it constituted the first initiative of its kind for invertebrates (though see Nilsson et al. 2018 for similar efforts in other taxa), acting as a model initiative for such activities in other taxonomic groups in the future. It also offered a first-of-its-kind human-mediated assessment of the taxonomic congruence status of publicly available barcode data in BOLD.

The investigation highlighted an important proportion of BOLD marine records that may lead to erroneous species identification, particularly those records tagged with MIS-ID and AMBIG (24% of reviewed species), in the context in which only a fraction (10%) of the world marine invertebrate species (of taxa of interest here) had any representation in BOLD. On the other hand, the revision also identified a relatively high proportion of species harboring undescribed intraspecific diversity (14% of species tagged as COMPLEX). Species with MIS-ID tags constituted a relatively small portion among the full set reviewed here (7%), although it is possible that some AMBIG records are MIS-ID but could not be fully resolved with the information available (see also discussion further below regarding AMBIG). The review uncovered substantial differences in the proportion of MIS-ID between taxonomic groups, with incidence percentages up to five to six times greater in Bivalvia and certain Gastropoda groups. This finding suggests that continued efforts to audit these two groups in particular are required. MIS-ID tags were below 4% in the remaining taxonomic groups. Despite the fact that misidentifications are not a very concerning fraction of the records, depending on the taxonomy and context of the research where the data is used (e.g., detection of non-indigenous marine species, see Bortolus 2008), attempts to detect them should continue. More than 2,700 records (over 500 species) of marine invertebrates were flagged and removed from BOLD IDS as a result of the hackathon, avoiding mistakes from being perpetuated in future research that rely only on a molecular identification system. However, the hope is that most records will be corrected in time (e.g., based on taxonomic identification of voucher specimens) and the flags

removed. Whereas a taxonomic update can be conducted very easily by the data owner or the BOLD team for records originating in BOLD, a major issue concerns the records mined from GenBank into BOLD. Such records cannot be easily manipulated and were found to be an important source for BIN discordances during the hackathon. For instance, BOLD:AAC4712 was composed of 27 records identified as *Talitrus saltator* and originating from three research labs in Canada, Germany and Portugal, hence identified by different specialists, and one record mined from GenBank and identified as *Talorchestia* sp. 1 HL-2018. Regardless of the interim species name status, in itself a source of BIN discordances, the record was deemed to be MIS-ID due to the genus level mismatch, hence flagged by the BOLD team and removed from BOLD IDS. In this case, the probability of a correction and removal of the flag is slim since it would need a complex operation in two databases (morphological verification of the voucher specimen, taxonomic correction in GenBank by one of the authors of the study generating that sequence, followed by taxonomic correction of the mined record by the BOLD team).

The fact that the majority of AMBIG tags were applied to uncertain data was not surprising because the review took a conservative approach, and this tag was used as a last resort when no other tag could be reliably assigned. This might have inflated the number of AMBIG tags that would have been assigned to other categories if this precautionary approach had not been taken, but it is impossible to ascertain to what extent. On the other hand, a detailed taxa-partitioned inspection of the AMBIG records unraveled a highly unbalanced distribution, with some particular taxonomic groups like Nudibranchia, Littorinimorpha and Pulmonata contributing disproportionately to the global numbers of tagged species (26%, 31% and 54%, respectively; Suppl. material 1: Fig. S1), whereas in some other groups the proportion of AMBIG tags was comparatively low and even supplanted by other categories (e.g., Crustacea and Polychaeta). As a side note, while Pulmonata is no longer a valid taxon in WoRMS, it is still considered an order in BOLD, hence its use here. Littorinimorpha and Nudibranchia are particularly speciose taxa (e.g., 6,479 and 2,429 species respectively, WoRMS, June 1, 2019) with some unsorted taxonomic riddles, which underwent numerous taxonomic revisions resulting in a high number of synonymies (e.g., *Littorina obtusata* (Linnaeus, 1758) with more than 50 synonyms in WoRMS). Among the many *Littorina* species represented in BOLD, *Littorina aleutica* and *L. natica* were found in the same BIN (BOLD:ACB8366) and since no additional information was found regarding barcode sharing, the corresponding records were tagged as AMBIG. BOLD is performing routine taxonomic updates based on WoRMS taxonomy, therefore synonyms were not solved during this review to avoid interference in operational procedures.

It is important, but challenging, to discern between misleading data resulting from errors in the barcoding workflow, and inaccurate data resulting from a lack of basic taxonomic knowledge, unsolved taxonomic conundrums, unrecognized synonyms or a taxon's status being in flux. Some of the AMBIG tags may result from misidentifications, while others may simply indicate unsolved taxonomies that, if sorted out, may reveal congruence between molecular and non-molecular data. Eventually, part of the molecular data may even be evidencing the "true" species boundaries currently masked by complex morphological traits. AMBIG tagged records should therefore be taken as a signal for caution in their use unless the end-user can find additional information for their clarification. A potential solution would be to avoid species-level identification when using these tags, giving preference to higher rank assignments (although even errors at these ranks cannot be excluded with certainty). Recognition of taxonomic groups which have a large number of AMBIG tags could provide a focus for more detailed taxonomic work to clarify the status of various species.

The COMPLEX tag is the second most prevalent overall, but it is also the only one that does not necessarily preclude the accurate identification of specimens. It simply signals cases where possible undescribed intraspecific diversity was found. While usually COMPLEX meant a species split into two BINs, some cases of multiple splits were also found (e.g., *Capitella neoaciculata* with five BINs or *Paracorophium excavatum* with 15 associated BINs). Occurrences of multiple and highly divergent intraspecific lineages have been abundantly and increasingly reported in diverse groups of marine invertebrates, suggesting the existence of considerable hidden diversity (e.g., Nygren and Pleijel 2011, Lobo et al. 2017, Borges and Merckelbach 2018, Nygren et al. 2018, Desiderato et al. 2019, Teixeira et al. 2020). Several studies employed additional markers, mitochondrial and nuclear, essentially confirming the patterns observed with DNA barcodes (e.g., Borges and Merckelbach 2018, Hupało et al. 2019, Vieira et al. 2019). Despite the fact that most phyla and classes displayed values close to 10% or higher, the Crustacea (particularly Amphipoda) and Polychaeta had a higher and substantial proportion of species tagged with COMPLEX, even more than MIS-ID and SHARE (Fig. 4), indicating that this appears to be a common occurrence across the examined taxa. Curiously, the Gastropoda displayed comparatively low values with this tag. This may reflect in fact lower incidence of high-intraspecific divergences in this group but may also result in part from truly COMPLEX tags masked in the AMBIG category due to the high number of taxonomic discrepancies in the group.

Although not so critical for the accuracy of identifications, at least according to the current status of taxonomic knowledge, there are important aspects of the COMPLEX tag to consider. Most notably, it helps when perceiving the overall quantity of presumptive marine invertebrate species awaiting verification and eventual consolidation and description. Failing to recognize this considerable amount of hidden diversity may be just as detrimental for bioassessment and monitoring as the MIS-ID or AMBIG cases (Bickford et al. 2007). In general, very little is known about

potential biological and ecological differences among the highly divergent lineages, some of them confirmed species. Some of the species assigned with COMPLEX are prominent indicator species included in biotic indexes for bioassessment (e.g., AZTI Marine Biotic Index species list, Borja et al. 2000, https://ambi.azti.es/) and their overlooked intraspecific diversity may contribute to imprecisions and lower performance of the ecological status assessments.

A number of marine invertebrates with cosmopolitan or wide distributions are being discovered to be complexes comprising several units with narrower or restricted distributions (e.g., Gómez et al. 2007, Borges and Merckelbach 2018, Teixeira et al. 2020). Divergent lineages are frequently grouped together spatially, allowing genetic markers to be utilized to identify not just morphospecies but also regional or local lineages (e.g., Hupało et al. 2019, Vieira et al. 2019). The comprehensive detection of highly diverse intraspecific lineages is required for an accurate sense of changes in species ranges and occurrence, particularly in the scope of global change-induced responses. Those with smaller ranges are also more important in terms of biodiversity conservation (Bickford et al. 2007), as well as the consideration of marine protected zones and other conservation measures.

Species and records tagged with SHARE are by far the lowest proportion globally and within each taxonomic group. SHARE tags are associated with situations of low interspecific divergence coupled with incomplete sorting and haplotype sharing, as well as hybridization and introgression. As a result, rather than a reference library issue arising from flaws in the barcoding procedure, these indicate situations where the COI barcode sequences are unable to differentiate species based on values of genetic distances. They can, however, be used in situations of fully sorted and well-established species with records in the same BIN, sometimes separated by very low genetic distances. Either way, these results indicate that the occurrence of SHARE cases is minimal and can be promptly identified, or, in the latter case, circumvented through the accumulation of records into the libraries and refinement of the BIN assignment for that particular group. Gastropoda was again the group with the highest incidence of SHARE tags, reinforcing the perception that greater research effort is needed for taxonomic clarification of marine members of this group. As an example, *Littorina saxatilis* (BOLD:AAG1552) was found to share barcodes with *L. compressa* and *L. arcana*. Previous studies using various mitochondrial markers (NADH1, tRNApro, NADH6 and partial cytochrome *b* by Doellmann et al. 2011, COI by Borges et al. 2016) found *L. saxatilis* to be non-monophyletic with two mitochondrial lineages shared with the two sister species.

The BIN discordance report generated in BOLD is a highly valuable validation tool which easily highlights uncertain cases in need of careful examination. However, concordant BINs are not exempt from misidentification, especially less represented BINs, with barcodes from one project, thus probably identified by one person, or from multiple projects where BOLD users did not hold taxonomic information for their specimens and relied solely on the existing information in BOLD which, if erroneous initially, could have been propagated into their projects. In addition, singletons are very difficult, if not impossible, to verify. As the hackathon data included about 30–40% singletons for each taxonomic group investigated, it is possible that a larger proportion of the current marine data might need to be tagged with one of the four labels discussed above.

The challenges found in evaluating barcode data, particularly marine barcode data, point to the need for better practices when generating, analyzing, and publishing barcode data. BIN discordances owing to synonyms might be avoided with greater synchronization between WoRMS and BOLD. Interim species names, whether derived from original BOLD records or data mined from GenBank and accounting for a substantial proportion of BIN discordances, would benefit from being checked using BOLD IDS on a regular basis and updated if matches are found (although difficulties of taxonomic updates for GenBank-mined data have been already mentioned).

Conclusions

The one-day hackathon and the following months of annotation work contributed significantly to the curation of the BOLD DNA barcode reference libraries for key marine invertebrate groups. Although numerous significant taxonomic groups were omitted from analyses, it was still a massive undertaking that required the individual review and annotation of a large number of records and species. Despite the significant efforts, the hackathon only provided a snapshot of BOLD marine data from June 2019. Records that were flagged or tagged during the hackathon would ideally be cleared in a short period of time, by a coordinated effort by BOLD data owners together with the BOLD team, allowing them to be included in reliable and trustworthy barcode libraries.

Ideally, this kind of event should be repeated on a regular basis, in tandem with the addition of new entries to reference libraries. However, as a corollary of this enterprise, it was very evident that the immense effort required to complete this task cannot be underestimated, and that it could hardly be repeated in the same format.

Indeed, a much more practical approach is needed in future endeavors, and this pilot exercise provided some possible solutions to substantially simplify the review procedure. For instance, recent applications such as BAGS (Fontes et al. 2021), could help package data based on quality and prepare it for the review. On the other hand, a long-term solution would include the development of intelligent systems that can screen out the most obvious discordances and misleading records, and thus dispense human-mediated verification. The results presented here indicate that a considerable fraction of the discordances could benefit from such developments. The outstanding capabilities of machine learning (ML) and artificial intelligence (AI) systems have been extensively demonstrated

in various research fields (Davenport and Kalakota 2019) including in image-based species identification (e.g., Ärje et al. 2020, Høye et al. 2021) and in the analysis of metabarcoding data for ecological status assessment (e.g., Cordier et al. 2019, Frühe et al. 2020). Nonetheless, it was also evident from our results that there would always be a fraction of discordances that cannot be addressed through automated systems and would thus require human intervention to be properly annotated.

Therefore, whereas ML and AI-type of approaches may help to considerably reduce the number of records requiring review, turning hackathon-like initiatives into practical and feasible commitments, at the end of the line there will be the need for human-mediated verification at least, and hopefully, for a minor set of records. In this regard, DNA barcode reference libraries are no different from other biodiversity data, and, ideally, strategies for data curation through community involvement, similar to the community of editors curating taxonomic data on WoRMS, could be used as inspiration and transposed to the DNA barcoding practice.

Acknowledgements

The hackathon was organized with financial support from the European Union COST Action DNAqua-Net (CA 15219 https://dnaqua.net/) in the scope of the 8th International Barcode of Life Conference in Trondheim, Norway on 16 June 2019. DNAqua-Net is acknowledged for the funding provided and the local conference organizers for all the logistical support that ensured a successful event. Tyler Elliot and the rest of the BOLD team are acknowledged for their help with data queries and analytics. The authors also thank the hackathon participants for vibrant discussions during and after the event: Berry van der Hoorn, Katrine Konsghavn, Guy Paz, Mouna Rifi, Malin Strand, Anne Helene Tandberg, Adam Wall, and Endre Willassen. Marcos A. L. Teixeira was supported by a PhD grant from the Portuguese Foundation for Science and Technology (FCT I.P.) co-financed by ESF (SFRH/BD/131527/2017). Financial support granted by FCT to Sofia Duarte (CEECIND/00667/2017) and to Pedro E. Vieira (project NISDNA, PTDC/BIA-BMA/29754/2017) is also acknowledged. Sanna Majaneva was financially supported by the Norwegian Taxonomy Initiative (project no. 70184235). The authors thank the five reviewers who provided valuable input into the earlier version of the manuscript.

References

Andújar C, Creedy TJ, Arribas P, López H, Salces-Castellano A, Pérez-Delgado AJ, Vogler AP, Emerson BC (2021) Validated removal of nuclear pseudogenes and sequencing artefacts from mitochondrial metabarcode data. Molecular Ecology Resources 21: 1772–1787. https://doi.org/10.1111/1755-0998.13337

Ärje J, Raitoharju J, Iosifidis A, Tirronen V, Meissner K, Gabbouj M, Kiranyaz S, Kärkkäinen S (2020) Human experts vs. machines in taxa recognition. Signal Processing: Image Communication 87: 115917. https://doi.org/10.1016/j.image.2020.115917

Backeljau T (2018) *Crassostrea gigas* or *Magallana gigas*: A community-based scientific response. National Shellfisheries Association Quarterly Newsletter: 3.

Bayne BL, Ahrens M, Allen SK, D'Auriac MA, Backeljau T, Beninger P, Bohn R, Boudry P, Davis J, Green T, Guo X, Hedgecock D, Ibarra A, Kingsley-Smith P, Krause M, Langdon C, Lapègue S, Li C, Manahan D, Mann R, Perez-Paralle L, Powell EN, Rawson PD, Speiser D, Sanchez JL, Shumway S, Wang H (2017) The proposed dropping of the genus *Crassostrea* for all Pacific cupped oysters and its replacement by a new genus *Magallana*: A dissenting view. Journal of Shellfish Research 36: 545–547. https://doi.org/10.2983/035.036.0301

Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I (2007) Cryptic species as a window on diversity and conservation. Trends in Ecology and Evolution 22: 148–155. https://doi.org/10.1016/j.tree.2006.11.004

Bidartondo MI (2008) Preserving accuracy in GenBank. Science 319: 1616a–1616a. https://doi.org/10.1126/science.319.5870.1616a

Borges LMS, Hollatz C, Lobo J, Cunha AM, Vilela, AP, Calado G, Coelho R, Costa AC, Ferreira, MSG, Costa MH, Costa, FO (2016) With a little help from DNA barcoding: investigating the diversity of Gastropoda from the Portuguese coast. Scientific Reports 6: 20226. https://doi.org/10.1038/srep20226

Borges LMS, Merckelbach LM (2018) *Lyrodus mersinensis* sp. nov. (Bivalvia: Teredinidae) another cryptic species in the *Lyrodus pedicellatus* (Quatrefages, 1849) complex. Zootaxa 4442: 441–457. https://doi.org/10.11646/zootaxa.4442.3.6

Borja A, Franco J, Pérez V (2000) A marine Biotic Index to establish the ecological quality of soft-bottom benthos within European estuarine and coastal environments. Marine Pollution Bulletin 40: 1100–1114. https://doi.org/10.1016/S0025-326X(00)00061-8

Bortolus A (2008) Error cascades in the biological sciences: The unwanted consequences of using bad taxonomy in ecology. Ambio: A Journal of the Human Environment 37: 114–118. https://doi.org/10.1579/0044-7447(2008)37[114:ECITBS]2.0.CO;2

Buhay JE (2009) "COI-like" sequences are becoming problematic in molecular systematic and DNA barcoding studies. Journal of Crustacean Biology 29: 96–110. https://doi.org/10.1651/08-3020.1

Bucklin A, Peijnenburg KTCA, Kosobokova KN, O'Brien TD, Blanco-Bercial L, Cornils A, Falkenhaug T, Hopcroft RR, Hosia A, Laakmann S, Li C, Martell L, Questel JM, Wall-Palmer D, Wang M, Wiebe PH, Weydmann-Zwolicka A (2021) Toward a global reference database of COI barcodes for marine zooplankton. Marine Biology 168: e78. https://doi.org/10.1007/s00227-021-03887-y

Cordier T, Lanzén A, Apothéloz-Perret-Gentil L, Stoeck T, Pawlowski J (2019) Embracing environmental genomics and machine learning for routine biomonitoring. Trends in Microbiology 27: 387–397. https://doi.org/10.1016/j.tim.2018.10.012

Cristescu ME (2014) From barcoding single individuals to metabarcoding biological communities: Towards an integrative approach to the study of global biodiversity. Trends in Ecology and Evolution 29: 566–571. https://doi.org/10.1016/j.tree.2014.08.001

Davenport T, Kalakota R (2019) The potential for artificial intelligence in healthcare. Future Healthcare Journal 6: 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Desiderato A, Costa FO, Serejo CS, Abbiati M, Queiroga H, Vieira PE (2019) Macaronesian islands as promoters of diversification in amphipods: The remarkable case of the family Hyalidae (Crustacea, Amphipoda). Zoologica Scripta 48: 359–375. https://doi.org/10.1111/zsc.12339

Doellman MM, Trussell GC, Grahame JW, Vollmer SV (2011) Phylogeographic analysis reveals a deep lineage split within North Atlantic *Littorina saxatilis*. Proceedings of the Royal Society B: Biological Sciences 278: 3175–3183. https://doi.org/10.1098/rspb.2011.0346

Duarte S, Leite BR, Feio MJ, Costa FO, Filipe AF (2021) Integration of DNA-based approaches in aquatic ecological assessment using benthic macroinvertebrates. Water 13: 331. https://doi.org/10.3390/w13030331

Evans N, Paulay G (2012) DNA barcoding methods for invertebrates. In: Kress WJ, Erickson DL (Eds) DNA Barcodes. Methods in Molecular Biology (Methods and Protocols), vol 858. Humana Press, Totowa, 47–77. https://doi.org/10.1007/978-1-61779-591-6_4

Fontes JT, Vieira PE, Ekrem T, Soares P, Costa FO (2021) BAGS: An automated Barcode, Audit & Grade System for DNA barcode reference libraries. Molecular Ecology Resources 21: 573–583. https://doi.org/10.1111/1755-0998.13262

Frühe L, Cordier T, Dully V, Breiner H, Lentendu G, Pawlowski J, Martins C, Wilding TA, Stoeck T (2020) Supervised machine learning is superior to indicator value inference in monitoring the environmental impacts of salmon aquaculture using eDNA metabarcodes. Molecular Ecology 30: 2988–3006. https://doi.org/10.1111/mec.15434

Gómez A, Wright PJ, Lunt DH, Cancino JM, Carvalho GR, Hughes RN (2007) Mating trials validate the use of DNA barcoding to reveal cryptic speciation of a marine bryozoan taxon. Proceedings of the Royal Society B: Biological Sciences 274: 199–207. https://doi.org/10.1098/rspb.2006.3718

Grant DM, Brodnicke OB, Evankow AM, Ferreira AO, Fontes JT, Hansen AK, Jensen MR, Kalaycı TE, Leeper A, Patil SK, Prati S, Reunamo A, Roberts AJ, Shigdel R, Tyukosova V, Bendiksby M, Blaalid R, Costa FO, Hollingsworth PM, Stur E, Ekrem T (2021) The Future of DNA Barcoding: Reflections from Early Career Researchers. Diversity 13: 313. https://doi.org/10.3390/d13070313

Hobern D (2021) BIOSCAN: DNA barcoding to accelerate taxonomy and biogeography for conservation and sustainability. Genome 64: 161–164. https://doi.org/10.1139/gen-2020-0009

Høye TT, Ärje J, Bjerge K, Hansen OLP, Iosifidis A, Leese F, Mann HMR, Meissner K, Melvad C, Raitoharju J (2021) Deep learning and computer vision will transform entomology. Proceedings of the National Academy of Sciences of the United States of America 118: e2002545117. https://doi.org/10.1073/pnas.2002545117

Hupało K, Teixeira MAL, Rewicz T, Sezgin M, Iannilli V, Karaman GS, Grabowski M, Costa FO (2019) Persistence of phylogeographic footprints helps to understand cryptic diversity detected in two marine amphipods widespread in the Mediterranean basin. Molecular Phylogenetics and Evolution 132: 53–66. https://doi.org/10.1016/j.ympev.2018.11.013

Leese F, Altermatt F, Bouchez A, Ekrem T, Hering D, Meissner K, Mergen P, Pawlowski J, Piggott J, Rimet F, Steinke D, Taberlet P, Weigand A, Abarenkov K, Beja P, Bervoets L, Björnsdóttir S, Boets P, Boggero A, Bones A, Borja Á, Bruce K, Bursić V, Carlsson J, Čiampor F, Čiamporová-Zatovičová Z, Coissac E, Costa F, Costache M, Creer S, Csabai Z, Deiner K, DelValls Á, Drakare S, Duarte S, Eleršek T, Fazi S, Fišer C, Flot J-F, Fonseca V, Fontaneto D, Grabowski M, Graf W, Guðbrandsson J, Hellström M, Hershkovitz Y, Hollingsworth P, Japoshvili B, Jones J, Kahlert M, Kalamujic Stroil B, Kasapidis P, Kelly M, Kelly-Quinn M, Keskin E, Kõljalg U, Ljubešić Z, Maček I, Mächler E, Mahon A, Marečková M, Mejdandzic M, Mircheva G, Montagna M, Moritz C, Mulk V, Naumoski A, Navodaru I, Padisák J, Pálsson S, Panksep K, Penev L, Petrusek A, Pfannkuchen M, Primmer C, Rinkevich B, Rotter A, Schmidt-Kloiber A, Segurado P, Speksnijder A, Stoev P, Strand M, Šulčius S, Sundberg P, Traugott M, Tsigenopoulos C, Turon X, Valentini A, van der Hoorn B, Várbíró G, Vasquez Hadjilyra M, Viguri J, Vitonytė I, Vogler A, Vrålstad T, Wägele W, Wenne R, Winding A, Woodward G, Zegura B, Zimmermann J (2016) DNAqua-Net: Developing new genetic tools for bioassessment and monitoring of aquatic ecosystems in Europe. Research Ideas and Outcomes 2: e11321. https://doi.org/10.3897/rio.2.e11321

Leese F, Bouchez A, Abarenkov K, Altermatt F, Borja Á, Bruce K, Ekrem T, Ciampor Jr F, Ciamporová-Zaťovičová Z, Costa FO, Duarte S, Elbrecht V, Fontaneto D, Franc A, Geiger MF, Hering D, Kahlert M, Kalamuji Stroil B, Kelly M, Keskin E, Liska I, Mergen P, Meissner K, Pawlowski J, Penev L, Reyjol Y, Rotter A, Steinke D, van der Wal B, Vitecek S, Zimmermann J, Weigand AM (2018) Why we need sustainable networks bridging countries, disciplines, cultures and generations for Aquatic Biomonitoring 2.0: A perspective derived from the DNAqua-Net COST Action. Advances in Ecological Research 58: 63–99. https://doi.org/10.1016/bs.aecr.2018.01.001

Leite BR, Vieira PE, Teixeira MAL, Lobo-Arteaga J, Hollatz C, Borges LMS, Duarte S, Troncoso JS, Costa FO (2020) Gap-analysis and annotated reference library for supporting macroinvertebrate metabarcoding in Atlantic Iberia. Regional Studies in Marine Science 36: 101307. https://doi.org/10.1016/j.rsma.2020.101307

Lis JA, Lis B, Ziaja DJ (2016) In BOLD we trust? A commentary on the reliability of specimen identification for DNA barcoding: A case study on burrower bugs (Hemiptera: Heteroptera: Cydnidae). Zootaxa 4114: 83–86. https://doi.org/10.11646/zootaxa.4114.1.6

Lobo J, Ferreira MS, Antunes IC, Teixeira MAL, Borges LMS, Sousa R, Gomes PA, Costa MH, Cunha MR, Costa FO, Hogg I (2017) Contrasting morphological and DNA barcode-suggested species boundaries among shallow-water amphipod fauna from the southern European Atlantic coast. Genome 60: 147–157. https://doi.org/10.1139/gen-2016-0009

Mammola S, Riccardi N, Prié V, Correia R, Cardoso P, Lopes-Lima M, Sousa R (2020) Towards a taxonomically unbiased European Union biodiversity strategy for 2030. Proceedings of the Royal Society B: Biological Sciences 287: 20202166. https://doi.org/10.1098/rspb.2020.2166

McGee KM, Robinson CV, Hajibabaei M (2019) Gaps in DNA-based biomonitoring across the globe. Frontiers in Ecology and Evolution 7: 337. https://doi.org/10.3389/fevo.2019.00337

Meiklejohn KA, Damaso N, Robertson JM (2019) Assessment of BOLD and GenBank – Their accuracy and reliability for the identification of biological materials. PLoS ONE 14: e0217084. https://doi.org/10.1371/journal.pone.0217084

Nilsson RH, Taylor AFS, Adams RI, Baschien C, Bengtsson-Palme J, Cangren P, Coleine C, Daniel H-M, Glassman SI, Hirooka Y, Laszlo I, Iršėnaitė R, Martin-Sanchez PM, Meyer W, Oh S-Y, Sampaio JP, Seifert K, Sklenář F, Dirk Stubbe DS, Suh S-O, Summerbell R, Svantesson S, Unterseher M, Visagie C, Weiss M, Woudenberg J, Wurzbacher C, Van den Wyngaert S, Yilmaz N, Yurkov A, Kõljalg

U, Abarenkov K (2018) Taxonomic annotation of public fungal ITS sequences from the built environment – a report from an April 10–11, 2017 workshop (Aberdeen, UK). MycoKeys 28: 65–82. https://doi.org/10.3897/mycokeys.28.20887

Nygren A, Pleijel F (2011) From one to ten in a single stroke – resolving the European *Eumida sanguinea* (Phyllodocidae, Annelida) species complex. Molecular Phylogenetics and Evolution 58: 132–141. https://doi.org/10.1016/j.ympev.2010.10.010

Nygren A, Parapar J, Pons J, Meißner K, Bakken T, Kongsrud JA, Oug E, Gaeva D, Sikorski A, Johansen RA, Hutchings PA, Lavesque N, Capa M (2018) A mega-cryptic species complex hidden among one of the most common annelids in the North East Atlantic. PLoS ONE 13: e0198356. https://doi.org/10.1371/journal.pone.0198356

Padial JM, De La Riva I (2006) Taxonomic inflation and the stability of species lists: The perils of ostrich's behavior. Systematic Biology 55: 859–867. https://doi.org/10.1080/1063515060081588

Padial JM, De la Riva I (2020) A paradigm shift in our view of species drives current trends in biological classification. Biological Reviews 96: 731–751. https://doi.org/10.1111/brv.12676

Patterson DJ, Cooper J, Kirk PM, Pyle RL, Remsen DP (2010) Names are key to the big new biology. Trends in Ecology and Evolution 25: 686–691. https://doi.org/10.1016/j.tree.2010.09.004

Pentinsaari M, Ratnasingham S, Miller SE, Hebert PDN (2020) BOLD and GenBank revisited – Do identification errors arise in the lab or in the sequence libraries? PLoS ONE 15: e0231814. https://doi.org/10.1371/journal.pone.0231814

Porter TM, Hajibabaei M (2018) Over 2.5 million COI sequences in GenBank and growing. PLoS ONE 13: e0200177. https://doi.org/10.1371/journal.pone.0200177

Radulovici AE, Archambault P, Dufresne F (2010) DNA barcodes for marine biodiversity: Moving fast forward? Diversity 2: 450–472. https://doi.org/10.3390/d2040450

Ramirez JL, Rosas-Puchuri U, Cañedo RM, Alfaro-Shigueto J, Ayon P, Zelada-Mázmela E, Siccha-Ramirez R, Velez-Zuazo X (2020) DNA barcoding in the Southeast Pacific marine realm: Low coverage and geographic representation despite high diversity. PLoS ONE 15: e0244323. https://doi.org/10.1371/journal.pone.0244323

Ratnasingham S, Hebert PDN (2007) BOLD: The Barcode of Life Data System (www.barcodinglife.org). Molecular Ecology Resources 7: 355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x

Ratnasingham S, Hebert PDN (2013) A DNA-based registry for all animal species: The Barcode Index Number (BIN) System. PLoS ONE 8: e66213. https://doi.org/10.1371/journal.pone.0066213

Rimet F, Aylagas E, Borja Á, Bouchez A, Canino A, Chauvin C, Chonova T, Ciampor Jr F, Costa FO, Ferrari BJD, Gastineau R, Goulon C, Gugger M, Holzmann M, Jahn R, Kahlert M, Kusber W-H, Laplace-Treyture C, Leese F, Leliaert F, Mann DG, Marchand F, Méléder V, Pawlowski J, Rasconi S, Rivera S, Rougerie R, Schweizer M, Trobajo R, Vasselon V, Vivien R, Weigand A, Witkowski A, Zimmermann J, Ekrem T (2021) Metadata standards and practical guidelines for specimen and DNA curation when building barcode reference libraries for aquatic life. Metabarcoding and Metagenomics 5: 17–33. https://doi.org/10.3897/mbmg.5.58056

Salvi D, Macali A, Mariottini P (2014) Molecular phylogenetics and systematics of the bivalve family Ostreidae based on rRNA sequence-structure models and multilocus species tree. PLoS ONE 9: e108696. https://doi.org/10.1371/journal.pone.0108696

Sayers EW, Cavanaugh M, Clark K, Pruitt KD, Schoch CL, Sherry ST, Karsch-Mizrachi I (2021) GenBank. Nucleic Acids Research 49: D92–D96. https://doi.org/10.1093/nar/gkaa1023

Siddall ME, Fontanella FM, Watson SC, Kvist S, Erséus C (2009) Barcoding bamboozled by bacteria: Convergence to metazoan mitochondrial primer targets by marine microbes. Systematic Biology 58: 445–451. https://doi.org/10.1093/sysbio/syp033

Teixeira MAL, Vieira PE, Pleijel F, Sampieri BR, Ravara A, Costa FO, Nygren A (2020) Molecular and morphometric analyses identify new lineages within a large *Eumida* (Annelida) species complex. Zoologica Scripta 49: 222–235. https://doi.org/10.1111/zsc.12397

Vieira PE, Desiderato A, Holdich DM, Soares P, Creer S, Carvalho GR, Costa FO, Queiroga H (2019) Deep segregation in the open ocean: Macaronesia as an evolutionary hotspot for low dispersal marine invertebrates. Molecular Ecology 28: 1784–1800. https://doi.org/10.1111/mec.15052

Vieira PE, Lavrador AS, Parente MI, Costa AC, Costa FO, Duarte S (2021) Gaps in DNA sequence libraries for Macaronesian marine macroinvertebrates imply decades till completion and robust monitoring. Diversity and Distribution 27: 2003–2015. https://doi.org/10.1111/ddi.13305

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M, Szucsich N, Weigand AM, Willassen E, Wyler SA, Bouchez A, Borja A, Čiamporová-Zaťovičová Z, Ferreira S, Dijkstra KDB, Eisendle U, Freyhof J, Gadawski P, Graf W, Haegerbaeumer A, van der Hoorn BB, Japoshvili B, Keresztes L, Keskin E, Leese F, Macher JN, Mamos T, Paz G, Pešić V, Pfannkuchen DM, Pfannkuchen MA, Price BW, Rinkevich B, Teixeira MAL, Várbíró G, Ekrem T (2019) DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Science of the Total Environment 678: 499–524. https://doi.org/10.1016/j.scitotenv.2019.04.247

WoRMS Editorial Board (2019) World Register of Marine Species. http://www.marinespecies.org [Available at VLIZ, Accessed 2019-06-01]

Supplementary material 1
Figure S1 and Tables S1–S10

Author: Adriana E. Radulovici, Pedro E. Vieira, Sofia Duarte, Marcos A. L. Teixeira, Luisa M. S. Borges, Bruce E. Deagle, Sanna Majaneva, Niamh Redmond, Jessica A. Schultz, Filipe O. Costa

Data type: Image and tables (in zip. archive)

Explanation note: Figure S1. Number of species tagged with AMBIG within each order in the Gastropoda. Table S1. Amphipoda. Table S2. Bivalvia. Table S3. Gastropods1. Table S4. Gastropods2. Table S5. Gastropods3. Table S6. Gastropods4. Table S7. Gastropods5. Table S8. Crustacea. Table S9. Echinodermata. Table S10. Polychaeta.

Copyright notice: This dataset is made available under the Open Database License (http://opendatacommons.org/licenses/odbl/1.0/). The Open Database License (ODbL) is a license agreement intended to allow users to freely share, modify, and use this Dataset while maintaining this same freedom for others, provided that the original source and author(s) are credited.

Link: https://doi.org/10.3897/mbmg.5.67862.suppl1